Can you hear me now?
Towards talking Wordnets: A Cantonese Case Study

Joanna Ut-Seong Sio
Palacký University
joannautseong.sio@upol.cz

Luis Morgado da Costa
Vrije Universiteit Amsterdam
lmorgado.dacosta@gmail.com

Francis Bond
Palacký University
bond@ieee.org

Kamila Liedermannova
Palacký University
kamila.liedermannova01@upol.cz

Abstract

This paper describes an expansion of the Cantonese Wordnet with a special focus on hand-checked audio recordings containing the pronunciaton of Cantonese senses. To achieve this we explored an open dataset from Wikimedia Commons and also developed our own dataset targeting Cantonese L2 learners. We added audio recordings to 2,859 senses. This work also led to a considerable improvement in the coverage of this wordnet increasing its number of concepts by 18%.

1 Introduction

1.1 Intro to Cantonese Wordnet

The Cantonese Wordnet (Sio and Morgado da Costa, 2019) is a small but growing wordnet project for Cantonese.[1] It started in 2019 with a core set of synsets, and it has been expanding since. In 2022 it received the first update (Sio and Morgado da Costa, 2022), receiving an example corpus of of 3,570 example sentences, and expanding its coverage by both adding more senses and by moving beyond the structure of the Princeton Wordnet (Fellbaum, 1998) concept inventory to include Cantonese specific concepts such as classifiers. In 2023, the Cantonese Wordnet received another round of attention with the creation of the Open Cantonese Sense-Tagged Corpus (Sio and Morgado da Costa, 2023), a method known to help identify problems in the lexicon, such as missing or indistinguishable senses, while also contributing to attestation of language use (Miller et al., 1993).

1.2 Motivation

From its inception, the Cantonese Wordnet has been committed to being a high quality, hand-curated resource designed to support linguistic research, teaching and learning of Cantonese.

[1] https://github.com/lmorgadodacosta/CantoneseWN

The potential to use this wordnet for educational purposes was one of the main motivations to include hand-curated romanized transcriptions for each sense since its first release. The Cantonese Wordnet uses the Jyutping (粵拼) romanization system, developed by the Linguistic Society of Hong Kong (LSHK) in 1993 – which is widely used in Cantonese L2 Education. The creation of the Open Cantonese Sense-Tagged Corpus was also partially motivated by its usefulness to L2 learners of Cantonese – which can reliably get translations for all sense annotated words in a variety of languages using multilingual links provided by CILI – the Collaborative Interlingual Index (Bond et al., 2016), currently under development by the Global Wordnet Association.

This paper moves further in this direction, and follows through on previous plans to provide, alongside romanized transcriptions, audio recordings with pronunciations for individual senses. Audios of wordnet lemmas are an important addition, given that Cantonese is a tonal language predominantly used in speech. Furthermore, Cantonese has interesting phonetic variations. One example would be what commonly known by laymen as *laan5 jam1* "lazy pronunciations" (or more neutrally, Principle of Least Effort, or Principle of (Maximum) Ease of Articulation) (Ladefoged and Johnson, 2006; Bauer, 2016). Some examples of "lazy pronunciations" include /n/ and /l/ alternation (alveolar nasal alternates with lateral approximant, e.g., *naam4* and *laam4* (both mean "male")), or /ng/ and /∅/ alternation (velar nasal alternates with zero initial, e.g., *ngaa4* and *aa4* (both mean "tooth")) (Bauer, 2016). We include phonetic variations in our audio files whenever possible.

In its role as a resource to support teaching and learning of Cantonese, we believe audio recordings to be quite important. This is also evidenced by the fact that many learner dictionaries (in many languages) often include pronunciations. Specif-

ically for Cantonese, e.g., CantoWords[2] provides such recordings. And while this is commendable and a great service to the Cantonese speaking and learning community, dictionaries like these have their limitations. An important limitation concerns its license. While less restrictive than others, its non-commercial license limits its use and integration in other research projects with a more permissible license (as is the case for the Cantonese Wordnet). Another problem is its quality. For words with multiple characters, CantoWords provides audio recordings that are the concatenation of multiple recordings. This is most certainly better than no recordings, but this type of recording is not able to capture the pronunciation, prosody and tonal nuances of complex words (e.g., tone sandhi).

We believe Cantonese should have an open dataset of high-quality recordings. Such a dataset would be a valuable resource to support linguistic research, the learning community, but also the development of Cantonese tools (including commercial ones).

2 Extending the Cantonese Wordnet

The work presented in this paper focuses on two complementary datasets: i) Cantonese audio extracted from Wikimedia Commons[3] and ii) a new dataset created for Cantonese L2 education.

2.1 Wikimedia Commons Data

We were first made aware of this dataset when we were contacted by Wikimédia France[4] to raise awareness of Lingua Libre[5] – an online platform that allows users to preserve linguistic diversity by allowing users to record words, phrases and proverbs in their own language. This tool is integrated with Wikimedia Commons, which is where the audio recordings are stored.

There was one particular user (Luilui6666)[6] who was actively helping record Cantonese audio, and who was happy to help the Cantonese Wordnet project. In total, this user shared 6,198 audio recordings, 1,645 of them explicitly marked as Cantonese (i.e., by using the language code 'yue'). Unfortunately, the information provided about these files includes only simplified Chinese characters. This is problematic for two reasons: i) Cantonese in Hong Kong uses traditional characters (using simplified characters likely hinders their discoverability); and ii) the conversion between simplified characters is not lossless. A classic example is the conversion of 发 *faat3* meaning "to send" or "hair" , the former is expressed by the traditional form 發 and the latter 髮.

Status	No. of Recordings
Rejected	297
Failed Links	57
Linked to New Senses	648
Linked to Existing Senses	643
Total	1,645

Table 1: Wikimedia Commons Data Summary

Since the Cantonese Wordnet is still a small project, we knew that many of these recordings could include words still missing from the wordnet. We proceeded to convert the file names from simplified to traditional characters (so they could be matched with the wordnet).[7] The full list of recordings was then matched against existing senses based on their written form. Out of the total number of recordings, 643 were found in the Cantonese wordnet, and 1002 were missing.

We proceeded to automatically produce Jyutping transcriptions for each missing lemma[8] and to link them by hand to the wordnet when it was possible and desirable. We found the data contained three main choices: i) the lemma was rejected (i.e., either was not a single lemma or were not strictly Cantonese); ii) the lemma was deemed fit to be added to the wordnet but we were not able to do so (this can happen for a variety of reasons, which will be discussed in greater detail below); and iii) the lemma was added to the Cantonese wordnet. Table 1 shows a summary of the results.

2.2 Cantonese L2 Data

As a complement to the Wikimedia Commons Data, we also started collecting our own recordings. We employed the help of a Cantonese L2 lecturer and native speaker to create a list of common words used in Cantonese L2 teaching. The same individual recorded high quality readings of this list using professional grade equipment. If a word had more

[2] https://cantowords.com/
[3] https://commons.wikimedia.org
[4] https://www.wikimedia.fr/
[5] https://lingualibre.org/
[6] https://commons.wikimedia.org/wiki/User:Luilui6666
[7] https://pypi.org/project/chinese-converter
[8] https://pypi.org/project/pinyin-jyutping

Status	No. of Recordings
Failed Link	112
Linked to New Senses	349
Linked to Existing Senses	406
Total	867

Table 2: Cantonese L2 Data Summary

than one possible pronunciation, multiple recordings were made. This generated 867 recordings.

We followed the same procedure described above. The 867 recordings were matched against the Cantonese Wordnet linking 406 recordings to existing senses. Of the 461 remaining recordings, 349 were hand-linked to the wordnet with new links, and 112 failed to be linked (see discussion below). A summary is shown in Table 2.

3 Discussion

3.1 Wikimedia Commons Data

There are a few hurdles in incorporating all the Wikimedia Commons audio recordings into the Cantonese Wordnet. The main problem is that not all of the items are Cantonese, but rather Hong Kong "Chinese", a distinction that requires some discussion.

All literate Cantonese speakers use two different varieties of Chinese in Hong Kong. Cantonese is used at home, in everyday conversation and to a more limited extent, in informal writing (e.g., social media, tabloid magazines, etc.). Cantonese is not taught in school. Standard written Chinese (Mandarin) is also used in Hong Kong. It is taught in schools and is used in formal writings. This is Hong Kong "Chinese". Hong Kong "Chinese" is generally not spoken, except in cases like poetry recital or reading out loud from books.

All standard written Chinese lexical items can be pronounced in Cantonese. For example, the lemma for "umbrella" in standard Chinese is 雨傘, pronounced as *yǔsǎn* in Mandarin. The term is not used in Cantonese, but can be pronounced in Cantonese as *jyu5 saan3*. [9] Furthermore there is a substantial overlap between the lexical items in standard written Chinese and Cantonese. As a result, it is not always easy to distinguish the two varieties (Bauer, 2016). The Cantonese Wordnet, as of now, only contains items that belong to Cantonese as a spoken variety. The rule of thumb is that we only include items that we think would appear in spoken Cantonese (including both informal conversation and more formal contexts, such as news reports).

The Wikimedia Commons recordings contain mostly (possibly only) items that are used in standard Chinese (Mandarin). The recordings are items pronounced in Cantonese. Many of these items are also used in Cantonese, but not all. Thus, not all of the items are usable (e.g., 沸騰 *fèiténg* "to boil" is used in Mandarin but would instead be expressed as 滾 *gwan2* in Cantonese). Furthermore, the recordings do not only contain single lemmas but also phrases (e.g., 在…時候 " during...the time") and sentences (e.g., 他的褲子和他的襯衣不相配 "His pants and his shirt don't match"). These are not entries suitable for the Cantonese Wordnet, as they can be decomposed further. At any rate, the audio recordings of the longer chunks, phrases and sentences are all standard Chinese (Mandarin) pronounced in Cantonese. These phrases and sentences contain non-Cantonese constructions and lexical items, e.g., the Mandarin 的, *de* "of", which should be 嘅 *ge3* in Cantonese; the Mandarin 在 *zài*, "be at", which should be 喺 *hai2* in Cantonese; the Mandarin 喝, *hē*,"drink", which should be 飲, *jam2*, in Cantonese. The audio recordings of these longer phrases also seem to be put together with smaller chunks in such a way that the sound segments overlap, rendering them unintelligible. Despite these challenges, the Wikimedia Commons dataset offers a valuable resource for expanding Cantonese word coverage, provided careful attention is paid to the distinction between Hong Kong "Chinese" and spoken Cantonese usage, and checking of the individual audio files.

Among lemmas (in the recordings) that we would like to include in the Cantonese wordnet but do not yet exist in the Princeton WordNet, they are mainly idioms (e.g., 供不應求 *gung1 bat1 ying3 kau4* "demand is greater than supply"), which do not correspond to single lemmas, and culturally (Chinese) specific items (e.g., 繁體字 *faan4 tai2 zi6* "(Chinese) traditional characters", 拜年 *baai3 nin4* "going to visit relatives during Chinese New Year holidays").

3.2 Cantonese L2 data

The second set of audio recordings was recorded and compiled by a native Cantonese speaker (who was also a lecturer of Cantonese). There was no problem with the Cantonese-ness of the data. However, we still have to exclude some of the data for

[9] The word for umbrella in Cantonese is 遮 *ze1*.

now because they are arguably phrases. For example, 煲劇 *bou1 kek6* "to binge-watch a series", which contains both the verb 煲 *bou1* "to cook/boil" and the noun 劇 *kek6* "TV series". The syntactic status of these V-O sequences (compound words vs. phrases) are controversial in Cantonese grammar because on the one hand, they allow the insertion of aspectual particles in-between the two characters (V and O), e.g., the progressing aspectual particle 緊 *gan2* can be added in the middle of 煲劇 *bou1 kek6*: 煲緊劇 *bou1 gan2 kek6*, meaning the binge-watching of some series is in progress. On the other hand, their combinations are partially idiomatic in nature. For 煲劇 *bou1 kek6*, the verb 煲 *bou1* means "to cook/boil", not "watch". It only means "watch" in 煲劇 *bou1 kek6*. Furthermore, the metaphorical usage of the verb 煲 *bou1* "to cook/boil" is semi-productive. It can be used in a limited number of phrases like 煲電話粥 *bou1 din6 waa2 zuk1*, literally it means "to boil telephone congee', which means "to talk on the phone for a very long time"). At this moment, we have decided to not include such examples in the Cantonese Wordnet, though these are items that would be very useful for L2 learners. In the future, we would like to explore using two level of annotation for these idiomatic but compositional items, mapping them to both multi-word expressions and also decompositionally (Sio and Morgado da Costa, 2023).

Among lemmas (in the recordings) that we would like to include in the Cantonese wordnet but which do not have appropriate concepts in the Princeton WordNet, these include (i) locations in Hong Kong (e.g., 尖沙咀 *zim1 saa1 zeoi2*, a waterfront area famous for shopping and views of Victoria Harbour), (ii) culturally (Hong Kong) specific items (e.g., 叮叮 *ding1 ding1*, Hong Kong trams) [10] and, (iii) Cantonese language specific phrases (e.g., 冚家鏟 *ham4 gaa1 caan2*, a swear word that literally means "whole family drop dead", meaning "bastard" or "jerk").

So far we have delayed adding new contentful concepts to the Cantonese Wordnet. In the future we will work with CILI (Bond et al., 2016) to create and share these concepts.

4 Release

Before the release of this paper, the Cantonese Wordnet had just over 5,250 concepts, lexified with around 16,300 senses. The work presented in this paper increased the sense inventory with 1050 new senses (695 new senses from the Wikimedia Commons dataset and 355 new senses from Cantonese L2 dataset). These new senses are distributed across 964 new concepts (665 from the Wikimedia Commons dataset and 308 from Cantonese L2 dataset) – an increase of 18%.

In addition to expanded coverage, the Cantonese Wordnet now also includes recordings for 2,138 unique pronunciations (combining both datasets), 15 of which have two sets of recordings (one male and one female voice). Each of these recordings corresponds to a hand checked (phonemic) Jyutping pronunciation.

Based on Jyutping matching , and given the occurrence of homophones, 1,809 senses previously included in the Cantonese Wordnet can now be linked to an audio recording. This is in addition to 1,050 new senses that were added as part of the work presented in this paper. In total, the Cantonese Wordnet currently has 2,859 senses linked to audio recordings.

4.1 Audio Recordings

The audio files for both datasets were originally encoded with Waveform Audio File Format (WAVE). We re-encoded the files using FLAC (Free Lossless Audio Codec),[11] which is lossless, so keeps all the information about the sound, and an open standard, therefore it is well supported. Because the snippets are single words, the audio files are sufficiently small (typically around 30kB).

These and all future audio recordings will become part of the standard release of the Cantonese Wordnet. Given their lightweight, the audio recordings will be released as part of the Cantonese Wordnet's original Github repository.[12]

4.2 Pronunciations in the Wordnet LMF

Similar to previous releases, this new version of the Cantonese Wordnet will be released in the Wordnet LMF format.[13]. Pronunciation was added to this LMF in McCrae et al. (2021, WN-LMF 1.1).

An example of its encoding is given in Figure 1. It uses the `<Pronunciation>` element. `<Pronunciation>` elements can currently be used as subelements of lemmas and forms. For the time

[10] The nickname 叮叮 *ding1 ding1* originates from the warning sounds of the bell, rung when the tram is in motion.

[11] https://xiph.org/flac/index.html

[12] https://github.com/lmorgadodacosta/CantoneseWN

[13] https://github.com/globalwordnet/schemas

```
<LexicalEntry id="cantown-yue-lex4">
    <Lemma writtenForm="你" partOfSpeech="n"/>
    <Form writtenForm="nei5">
        <Pronunciation variety="jyutping"
         audio ="https://path/nei5.flac">nei5</Pronunciation>
    </Form>
    <Form writtenForm="lei5">
        <Pronunciation variety="jyutping" notation="n/l lazy"
            audio ="https://path/lei5.flac">lei5</Pronunciation>
    </Form>
    </Lemma>
    <Sense id="cantown-yue-77000021-n-lex4"
    synset="cantown-yue-77000021-n"></Sense>
</LexicalEntry>
```

Figure 1: WN-LMF representation of pronunciation: the Cantonese lemma 你 has two forms *nei5* and *lei5*. Each form has a single pronunciation element (although multiple would be possible), linking it to the respective file.

being, we decided to keep it under form to allow clustering of multiple pronunciations by its Jyutping form. For example, in Figure 1, we currently have a single pronunciation for the Jyutping form *nei5*, but we see a future where we will have multiple recordings (e.g., using male and female voices, or representing regional variation). Another benefit of keeping Jyutping as a form, is the fact that many existing tools, such as the Open Multilingual Wordnet (Bond and Foster, 2013), display and even search over these fields. The `<Pronunciation>` element has 4 attributes:

- **audio**: provides the URL for the recording.
- **variety** encodes the language variety, for example by using the IETF language tags to indicate dialect, where Cantonese in jyutping would be `zh-yue-jyutping`. There is no general standard for how these are labelled, each wordnet can decide on its own — in our case `zh-yue` is redundant, so we omit it.
- **phonemic**: indicates whether the transcription is phonemic (**true**) or phonetic (**false**), with the default being **true** (phonemic).
- **notation**: can capture any additional details.

More detail of the audio support in the LMF is given in Bond (2025).

5 Future Work

The work presented in this paper is a step in a specific direction. As noted above, there are still many senses that do not have linked pronunciations. We would like to continue this work by exploring other open resources and continuing producing our own recordings to ensure that the majority of Cantonese senses, or at least those in common use, have a linked pronunciation. Of interest to this goal, is the fact that after inspecting the remainder of the data we extracted from Wikimedia Commons, it became clear many other recordings are pronounced in Hong Kong "Chinese". We believe that we can semi-automatically check which recordings may be of use to the Cantonese Wordnet – exploring homophones between Cantonese and Hong Kong "Chinese" to get additional recordings for the wordnet.

Another interesting avenue we would like to explore is exploring the new available recordings in the field of gamification of Cantonese learning and teaching. We believe the Cantonese Wordnet can soon be served as a learner's dictionary. And we would like to develop companion online apps to help Cantonese learners, gamifying the learning process of lexical and tonal knowledge.

6 Conclusion

This paper presents significant advancements in the Cantonese Wordnet, with the addition of 1,050 new senses across 964 new synsets and 2,138 new audio recordings, substantially improving its utility for both linguistic research and L2 Cantonese education. By incorporating pronunciations and linking them to specific senses, we have made the resource more accessible, especially in the context of Cantonese's tonal phonology. Moving forward, we plan to expand the dataset further, incorporat-

ing more educational and culturally specific lemmas, and explore ways to use this data in language-learning applications.

Acknowledgments

We would like to thank WikiCommons user Luilui6666 for changing the license of her recordings to accommodate the Cantonese Wordnet's license. We would also like to thank Dennis Lam for recording the the list of Cantonese L2 common lexical items for us.

Kamila Liedermannova would like to thank the IGA student grant funded by Palacký University (IGA_FF_2033_063 SPP: 432105081/31) for supporting this research study.

References

Robert S Bauer. 2016. The hong kong cantonese language: Current features and future prospects. *Global Chinese*, 2(2):115–161.

Francis Bond. 2025. Adding audio to wordnets,. In *13th International Global Wordnet Conference (GWC 2025)*. (this volume).

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Francis Bond, Piek Vossen, John Philip McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Peter Ladefoged and Keith Johnson. 2006. A course in phonetics (5th). *Thomson Wadsworth*, pages 133–236.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luís Morgado da Costa. 2021. The global wordnet formats: Updates for 2020. In *11th International Global Wordnet Conference (GWC2021)*.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Joanna Sio and Luis Morgado da Costa. 2023. The open cantonese sense-tagged corpus. In *Proceedings of the 12th Global Wordnet Conference*, pages 263–268.

Joanna Ut-Seong Sio and Luis Morgado da Costa. 2019. Building the cantonese wordnet. In *Proceedings of the 10th Global Wordnet Conference (GWC 2019), Wroclaw, Poland*.

Joanna Ut-Seong Sio and Luis Morgado da Costa. 2022. Enriching linguistic representation in the cantonese wordnet and building the new cantonese wordnet corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, Marseille, France. European Language Resources Association (ELRA).